More Confidence Intervals: Two Small Samples and Paired Data

Keegan Korthauer

Department of Statistics

UW Madison

RECAP: TWO SAMPLE CONFIDENCE INTERVALS

Large sample difference in population means

Difference in population proportion

CI for $\mu_X - \mu_Y$ with Large Samples

- Two large, **independent** samples: $n_X \geq 30$ and $n_Y \geq 30$
- Population means: μ_X and μ_Y
- Population variances: σ^2_X and σ^2_Y
- Level 100(1-α)% CI for the difference in means $\mu_X - \mu_Y$:

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

- When σ^2_X and σ^2_Y are unknown, replace them with the sample standard deviations: s^2_X and s^2_Y

CI for $p_X - p_Y$ with Two Samples

- Two populations with success probabilities: p_X and p_Y
- Two random samples containing X and Y successes out of n_X and n_Y
- Level 100(1-α)% CI for the difference in proportions $p_X - p_Y$:

$$(\tilde{p}_X - \tilde{p}_Y) \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}_X(1-\tilde{p}_X)}{\tilde{n}_X} + \frac{\tilde{p}_Y(1-\tilde{p}_Y)}{\tilde{n}_Y}}$$

$$\text{where } \tilde{n}_X = n_X + 2,\ \tilde{n}_Y = n_Y + 2,$$

$$\tilde{p}_X = (X+1)/\tilde{n}_X,\ \tilde{p}_Y = (Y+1)/\tilde{n}_Y$$

- $n_X > 4$ and $n_Y > 4$
- Truncate to [-1,1]

CI FOR THE DIFFERENCE IN TWO POPULATION MEANS (SMALL SAMPLE)

Difference in Means for Small Samples

- When both samples are small ($n_X < 30$ and $n_Y < 30$), the CLT does not apply so we can't apply the method for constructing large-sample CIs

- If we know that the two population distributions are approximately normal, we can make use of the Student's *t* distribution
  - recall that s is not a good estimate for σ when the sample size is small, so the *t* distribution accounts for the additional uncertainty with fatter tails

Recall the Student's *t* Distribution

Let $X_1,\ldots,X_n$ be a **small** ($n < 30$) sample from a **normal** population with mean μ. Then

$$\frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$

where t_{n-1} is the Student's *t* distribution with *n-1* degrees of freedom

The distribution is always centered at zero and has only one parameter (degrees of freedom *n-1*) that determines its shape

As *n* gets very large, t_{n-1} approaches N(0, 1)

Apply Student's *t* to Two Samples

Let $X_1,\ldots,X_{nx}$ and $Y_1,\ldots,Y_{ny}$ be **small** ($n_X < 30$ and $n_Y < 30$) samples from two **independent normal** populations with means μ_X and μ_Y. Then

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\dfrac{s_X^2}{n_X}+\dfrac{s_Y^2}{n_Y}}} \sim t_\nu$$

But, the degrees of freedom are not so straightforward with two samples:

$$\nu = \frac{\left(\dfrac{s_X^2}{n_X}+\dfrac{s_Y^2}{n_Y}\right)^2}{\dfrac{(s_X^2/n_X)^2}{n_X-1}+\dfrac{(s_Y^2/n_Y)^2}{n_Y-1}}$$

rounded **down** to the nearest integer

CI for $\mu_X - \mu_Y$ with Small Samples

- Let $X_1, \ldots, X_{nx}$ and $Y_1, \ldots, Y_{ny}$ be **small** ($n_X < 30$ and $n_Y < 30$) samples from two **normal** populations with means μ_X and μ_Y
- If the samples are **independent** and the population variances are not necessarily equal, then a Level 100(1-α)% CI for the difference in means $\mu_X - \mu_Y$ is:

$$(\bar{X} - \bar{Y}) \pm t_{\nu,\, \alpha/2} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

- Where the degrees of freedom (ν) are calculated using the formula on the previous slide (rounded down to the nearest integer)

Example – Treatment vs Control

6 subjects were given a drug (Treatment group) and an additional 6 subjects a placebo (Control group). Their reaction time to a stimulus was measured (in ms).

The sample mean of the treatment group was 83.83 ms (sd 7.17)

The sample mean of the control group was 101.67 ms (sd 5.65)

Let μ_T be the mean of the treatment population and μ_C the mean of the control population. We want to find a 95% CI for the difference in means of the treatment and control groups $\mu_T - \mu_C$.

Assume the two groups are independent, and that the treatment and control populations have an approximate normal distribution.

Special Case – Equal Population Variances

- If the population variances are known to be equal ($\sigma^2_X = \sigma^2_Y$), we can simplify the expression for the confidence interval
- Note that we can do this even if the actual value of $\sigma^2 = \sigma^2_X = \sigma^2_Y$ is unknown
- When we assume a common population variance σ^2 we can estimate the **Pooled standard deviation s_p**:

$$s_p = \sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}}$$

- This estimate of σ^2 uses both samples

Derive the Pooled Variance Estimate

- When we assume X and Y have equal variance, then

$$\bar{X} \sim N(\mu_X, \sigma^2 / n_X) \text{ and } \bar{Y} \sim N(\mu_Y, \sigma^2 / n_Y)$$

- Then it follows that

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\right)$$

- Since σ^2 is unknown, we might estimate it with s^2_X or s^2_Y. But even better, we can estimate it using both samples with the **weighted average of s^2_X and s^2_Y**

$$\hat{\sigma}^2 = s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{(n_X - 1) + (n_Y - 1)}$$

CI for $\mu_X - \mu_Y$ with Small Samples (Special Equal Variances Case)

- Let $X_1,\ldots,X_{nx}$ and $Y_1,\ldots,Y_{ny}$ be **small** ($n_X < 30$ and $n_Y < 30$) samples from two **normal** populations with means μ_X and μ_Y
- If the samples are **independent** and the population variances are equal ($\sigma^2_X = \sigma^2_Y$), then a Level 100(1-α)% CI for the difference in means $\mu_X - \mu_Y$ is:

$$(\bar{X} - \bar{Y}) \pm t_{n_X + n_Y - 2,\ \alpha/2}\ s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$$

- Where s_p is the square root of the **pooled variance** estimate given on the previous slide
- Note that the degrees of freedom have a much simpler form

Example – New vs Old Textbook

- We want to compare average exam scores for students using a new textbook versus the old textbook
- The class is divided into two and randomly assigned one of the two books
- We assume that the population of exam scores will be approximately normal and that the **variances will be the same regardless of textbook version (use pooled estimate)**
- Using the following summary data, find a 99% CI for the mean difference in exam scores for the old compared to new text:

	Old Text	New Text
Mean	64.3	68.8
Sample SD	7.1	7.4
Sample size	21	23

Notes

- This 'equal variance' assumption is *very* strict
- This method can be quite unreliable when misused
- When the sample variances are nearly equal, it is tempting to assume the population variances are nearly equal as well
    - but remember that with small sample sizes, the sample variances may not approximate the population variances well
- The best practice is to **assume the variances are unequal unless it is quite certain that they are equal**

Paired Data

- So far, we've discussed methods for finding CIs of the difference in two means of **two independent samples**

- Sometimes we are interested in the difference in means of **two measurements made on the same sample**

- This type of data is called **paired data**

- Since paired observations made on the same sample are no longer independent, we need new methods to find CIs

Example – Gas Mileage

- A study was performed to test whether cars get better mileage on premium gas than on regular gas.
- Each of 10 cars was first filled with either regular or premium gas, decided by a coin toss, and the mileage for that tank was recorded.
- The mileage was recorded again **for the same cars** using the other kind of gasoline.
- The results:

Car	1	2	3	4	5	6	7	8	9	10
Premium	19	22	24	24	25	25	26	26	28	32
Regular	16	20	21	22	23	22	27	25	27	28
Difference	3	2	3	2	2	3	-1	1	1	4

Deriving the CI for Paired Data

- Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be the n **paired** observations
  - for example, X_i is the gas mileage on premium for the i^{th} car and Y_i is the gas mileage on regular for the i^{th} car
- Then let $D_i = X_i - Y_i$
  - for example, D_i is the difference in gas mileage for premium and regular gasoline for the i^{th} car
- We are interested in a CI for the difference in population means ($\mu_D = \mu_X - \mu_Y$)
- We can apply one-sample methods for constructing CIs to the sample of differences

CI for Differences of Paired Data

- Let $D_1,\ldots,D_n$ be a **small** random sample of $n \leq 30$ differences of pairs
- Assume the population of differences is approximately normal
- Let the (unknown) standard deviation of the differences be σ_D be estimated by the sample sd of the differences s_D
- Then a $100(1-\alpha)\%$ CI for the mean difference μ_D is given by

$$\bar{D} \pm t_{n-1,\ \alpha/2} \frac{s_D}{\sqrt{n}}$$

- If the sample size is **large** ($n > 30$), then a $100(1-\alpha)\%$ CI for the mean difference μ_D is given by

$$\bar{D} \pm z_{\alpha/2} \frac{s_D}{\sqrt{n}}$$

Notes About Paired Data

- Treating the sample of paired observations as a single sample of differences uses the methods we learned for constructing CIs for a population mean (see section 5.1 for large sample, and 5.3 for small sample)

- Using paired data instead of two independent samples provides an advantage when there is high variability within a single sample
  - in the gas mileage example, considering the observations as pairs makes the variability between the cars disappear
  - cars vary widely in their gas mileage, but almost all of them have higher gas mileage with premium compared to regular

Example – Gas Mileage

Car	1	2	3	4	5	6	7	8	9	10
Premium	19	22	24	24	25	25	26	26	28	32
Regular	16	20	21	22	23	22	27	25	27	28
Difference	3	2	3	2	2	3	-1	1	1	4

Construct a 95% CI for the mean difference in gas mileage when cars use premium versus regular gasoline (assuming the population of differences is approximately normal).

Next

- Prediction Intervals for single observations

- Chapter 6 – Hypothesis Testing